Karlsruhe Institute of Technology
Institute of Applied Informatics and Formal Description Methods (AIFB)
M.Sc. Tarek Saier
tarek.saier@kit.edu

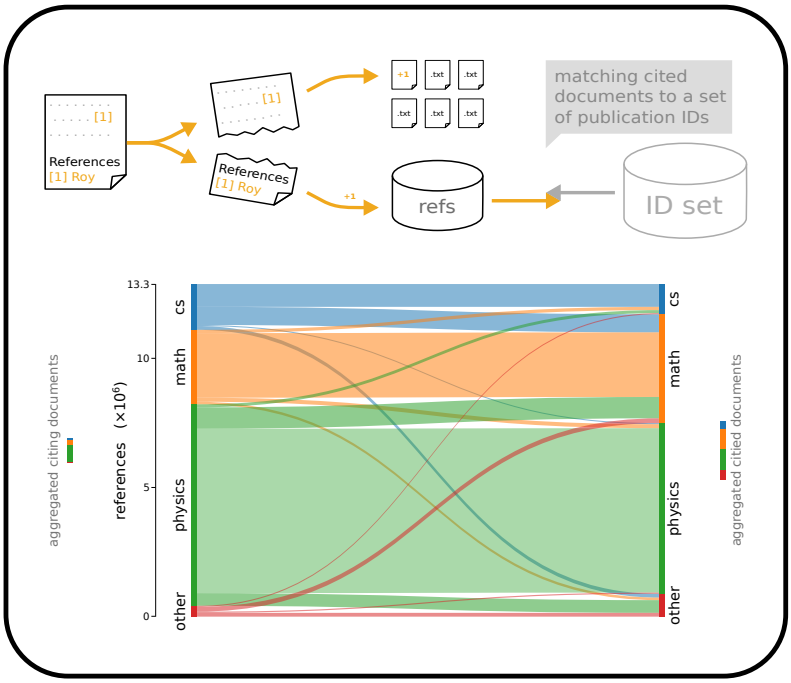
Call for Bachelor/Master Thesis
„Scholarly corpus creation - leveraging NLP to disambiguate bibliographic references“

What is the general goal?

The goal is to leverage Natural Language Processing and other Machine Learning techniques to build a robust method for linking bibliographic references from large corpora of academic publications to a common set of identifiers. Higher quality bibliographic links will facilitate improved information extraction from publications for e.g. trend detection and citation analysis.

What are the details?

With the growth of Open Access in academia and a need for automated processing of papers for knowledge extraction, the creation of machine readable scientific corpora is an active and evermore relevant area of research. However, current methods for disambiguating bibliographic references are still rather unsophisticated and lacking in performance.



Having developed our own corpus of over one million academic papers [1], we want to explore methods for a more sophisticated approach to disambiguate bibliographic references. Possible techniques to consider include NEL, blocking, end-to-end learning, zero-shot learning, etc.

What are the prerequesites?

- Interest in Data Mining, Machine Learning, and working with large data sets.
- No fear of actually implementing the methods you develop.

[1] Tarek Saier, Michael Färber, *"unarXive: A Large Scholarly Data Set with Publications' Full-Text, Annotated In-Text Citations, and Links to Metadata"*, Scientometrics. 2020.